MRC | Epidemiology Unit

Design and Analysis for EPIC-Norfolk GWAS of Obesity and Related Traits

Jing Hua Zhao

jinghua.zhao@mrc-epid.cam.ac.uk
http://www.mrc-epid.cam.ac.uk/~jinghua.zhao

11 June 2008, SRL

Table of contents

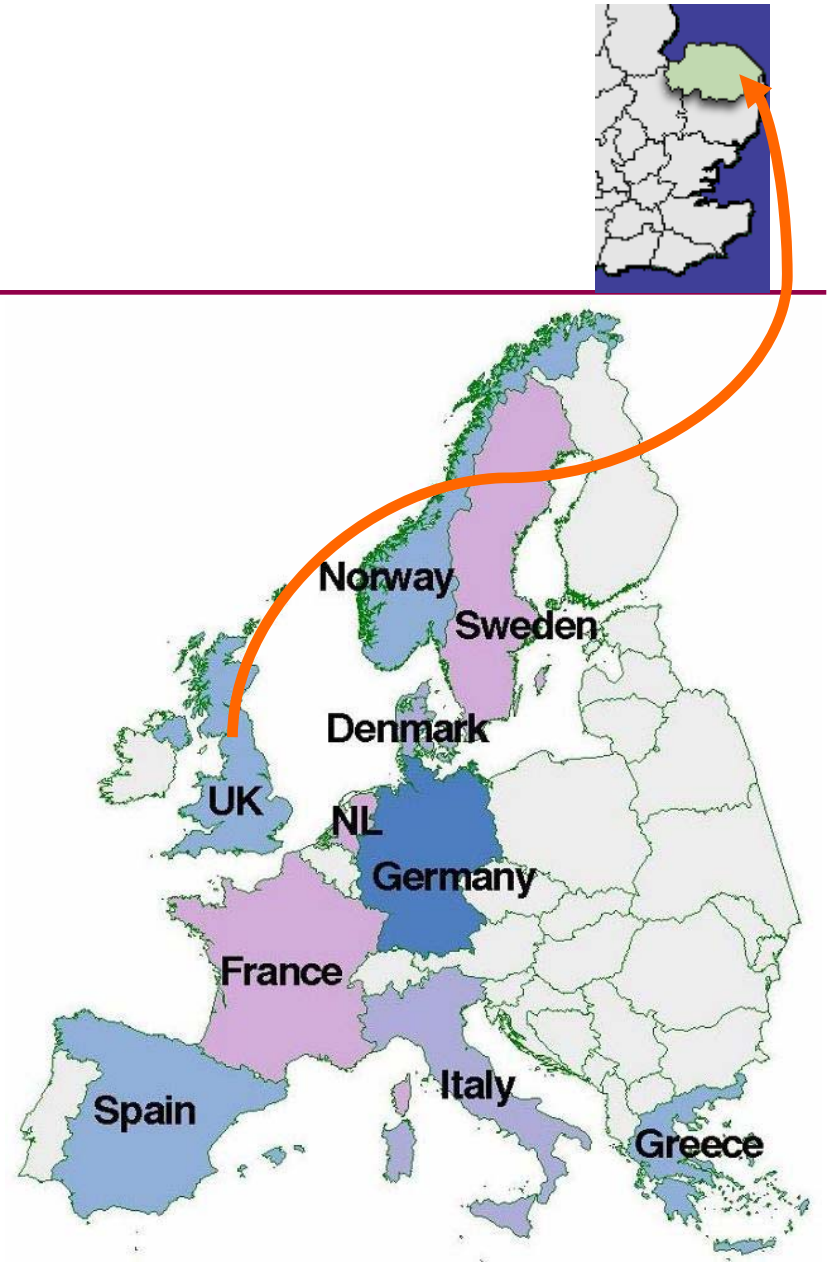
EPIC study

The European Prospective Investigation into Cancer and Nutrition (EPIC) is coordinated by Dr Elio Riboli, Head of the Division of Epidemiology, Public Health and Primary Care at the Imperial College London.

EPIC was designed to investigate the relationships between diet, nutritional status, lifestyle and environmental factors and the incidence of cancer and other chronic diseases. EPIC is the largest study of diet and health ever undertaken, having recruited over half a million (520,000) people in ten European countries: Denmark, France, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and the United Kingdom.

EPIC-Norfolk study

EPIC-Norfolk participants are men and women (based on over 30,000 people) who were aged between 45 and 74 when they joined the study, who lived in Norwich and the surrounding towns and rural areas. They have been contributing information about their diet, lifestyle and health through questionnaires, and through health checks carried out by EPIC nurses.



Designs for GWAS

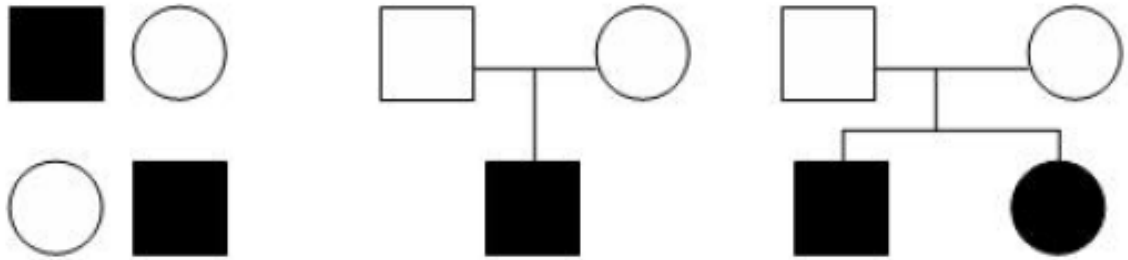


Figure 1: Three common genetic association designs involving unrelated individuals (left), nuclear families with affected singletons (middle) and affected sib-pairs (right). Males and females are denoted by squares and circles with affected individuals filled with black and unaffect individuals being empty.

Family designs for GWAS

		Linkage			Association					
γ	p	Y	N_L	P_A	H_1	N_{tdt}	H_2	$N_{asp/tdt}$	λ_o	λ_s
4.00	0.01	0.520	6400	0.800	0.048	1098	0.112	235	1.08	1.09
	0.10	0.597	277	0.800	0.346	151	0.537	48	1.48	1.54
	0.50	0.576	445	0.800	0.500	104	0.424	62	1.36	1.39
	0.80	0.529	3023	0.800	0.235	223	0.163	162	1.12	1.13
2.00	0.01	0.502	445839	0.667	0.029	5824	0.043	1970	1.01	1.01
	0.10	0.518	8085	0.667	0.245	696	0.323	265	1.07	1.08
	0.50	0.526	3752	0.667	0.500	340	0.474	180	1.11	1.11
	0.80	0.512	17904	0.667	0.267	640	0.217	394	1.05	1.05
1.50	0.01	0.501	6942837	0.600	0.025	19321	0.031	7777	1.00	1.00
	0.10	0.505	101898	0.600	0.214	2219	0.253	941	1.02	1.02
	0.50	0.510	27041	0.600	0.500	950	0.490	485	1.04	1.04
	0.80	0.505	101898	0.600	0.286	1663	0.253	941	1.02	1.02
Alzheimer's:										
4.50	0.15	0.626	163	0.818	0.460	100	0.621	37	1.67	1.78

Table 3: Comparison of linkage and association in nuclear families required for identification of disease gene: γ=genotypic risk ratio; p=frequency of disease allele A; Y=probability of allele sharing; N_L=number of ASP families required for linkage; P_A=probability of transmitting disease allele A; H_1, H_2=proportions of heterozygous parents; N_{tdt}=number of family trios; $N_{asp/tdt}$=number of ASP. families

Case-control designs for GWAS

		K			
γ	p	1%	5%	10%	20%
4.0	0.01	46638	8951	4240	1885
	0.10	8173	1569	743	331
	0.50	10881	2089	990	440
	0.80	31444	6035	2859	1271
2.0	0.01	403594	77458	36691	16307
	0.10	52660	10107	4788	2128
	0.50	35252	6766	3205	1425
	0.80	79317	15223	7211	3205
1.5	0.01	1598430	306770	145312	64583
	0.10	191926	36835	17448	7755
	0.50	97922	18793	8902	3957
	0.80	191926	36835	17448	7755

Table 5: Estimated sample sizes required for association detection using population data.

Case-cohort design for EPIC-Norfolk study

- It originally followed case-control design (e.g., WTCCC with seven cases and common controls) with 3425 cases and 3400 controls.
  - It is potentially more powerful.
  - Controls are selected.
- It has then been changed into case-cohort design, in which cases are defined to be individuals whose BMI above 30 and controls are a random sample (subcohort) of the EPIC-Norfolk cohort which includes obese individuals.
  - The subcohort is representative of the whole population and allows for a range of traits to be examined.
  - The analysis is potentially more involved but established.

Power/sample size

- It started with assessment of how the power is compromised relative to the original case-control design.
- This was followed by power/sample size calculation using methods established by Cai and Zeng (2004) as implemented in an R function, noting a number of assumptions.
- More practically, it was also envisaged that a proper representative sample of a total of 25,000 individuals would be 10%; the subcohort is then approximately 2,500.
- The total sample was split between two stages.

GeneChips

- Affymetrix 500K
  - Data were available for 3850 individuals
- Illumina 317K
  - It came at a later time。
  - Data quality appears to be poor?
- The focus has therefore been Affy500K, but with a possible comeback.

Analysis

- An incremental approach was adopted since the storage and computing power were somewhat uncertain.
- This was predated with controls from the breast cancer study, involving about 400 individuals with Perlegen 250K GeneChips.
- QC including call rates and HWE was feasible with SAS/Genetics (~30GB) which provides a good estimate of the storage for all individuals (~380GB).
- The Linux platform seemed to be favourable.

EPIC400 analysis

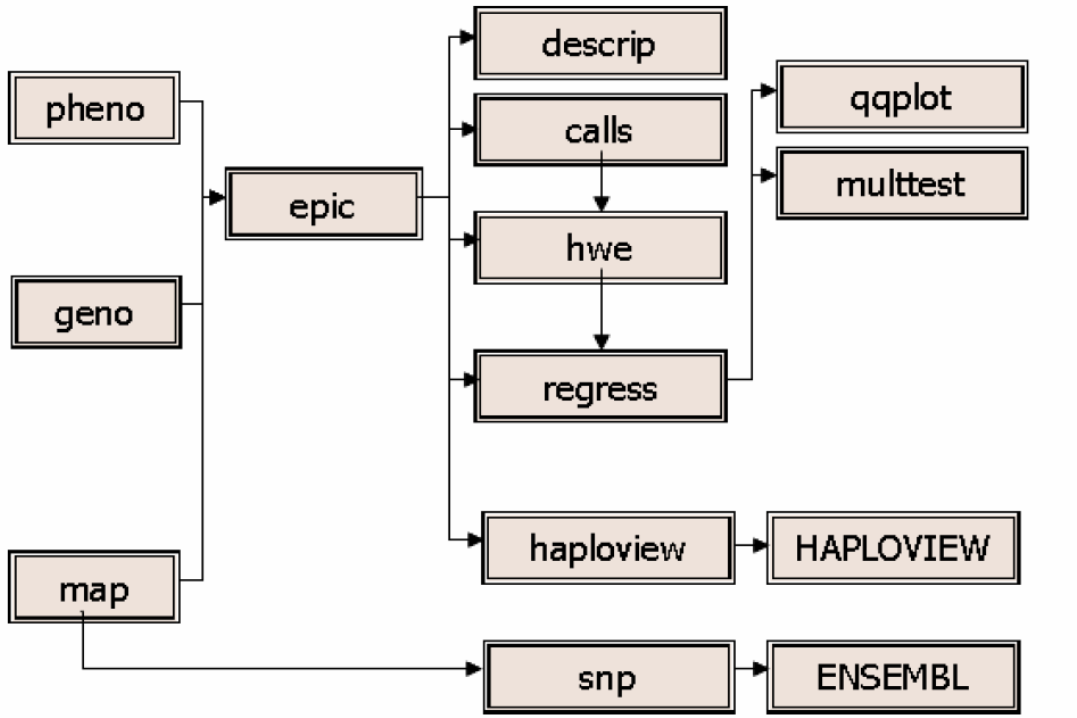


Fig. 1. A flowchart of the EPIC 400 analysis, with modules in brackets. Genotypes (geno) and phenotypes (pheno) are merged (epic) for descriptive statistics (descrip) call rates (calls), HWE (hwe), regression (regress) with adjustment for multiple testing (multtest) and comparison with theoretical distribution (qqplot). The raw data together with map information (map) can also be reformatted (haploview) into HAPLOVIEW input files so that specific region in the genome can be visualized, with annotation information from ENSEMBL according to SNPs (snp).

The analysis for GWAS

- QC including visualisation of clustering, outliers, was largely done by colleagues at Sanger (as for WTCCC)
- The overall strategy was data partition, i.e., by chromosome and further by region (30) in each chromosome, largely on a long, skinny data format
- A major advantage is that the analysis can be resumed whenever the system experiences problems
- We stuck to SAS to allow for reliability and flexibility with or without SAS/Genetics, for BMI/obesity as continuous and binary outcomes are readily tackled with REG/LOGISTIC procedures – most outputs are available from the output delivery system (ODS)
- The picture was eventually changed with a revised coding algorithm and the use of imputed data

Additional analysis

- Population stratification via EIGENSTRAT
  - SAS is very handy since a single put statement is sufficient to generate the output.
- Collaborative (e.g. height) and consortium work (GIANT)
  - On the UK side, this is mainly involved with IMPUTE/SNPTEST, with inputs on strand, standard error, quantitative traits, outputs.
  - This facilitates meta-analysis considerably.

LDL

LDL-cholesterol concentrations: a genome-wide association study

Manjinder S Sandhu, Dawn M Waterworth*, Sally L Debenham*, Eleanor Wheeler, Konstantinos Papadakis, Jing Hua Zhao, Kijoung Song, Xin Yuan, Toby Johnson, Sofie Ashford, Michael Inouye, Robert Luben, Matthew Sims, David Hadley, Wendy McArdle, Philip Barter, Y Antero Kesäniemi, Robert W Mahley, Ruth McPherson, Scott M Grundy, Wellcome Trust Case Control Consortium†, Sheila A Bingham, Kay-Tee Khaw, Ruth J F Loos, Gérard Waeber, Inês Barroso, David P Strachan, Panagiotis Deloukas, Peter Vollenweider, Nicholas J Wareham, Vincent Mooser*

BMI/obesity

Common variants near *MC4R* are associated with fat mass, weight and risk of obesity

Ruth J F Loos[*,1,2,73], Cecilia M Lindgren[3,4,73], Shengxu Li[1,2,73], Eleanor Wheeler[5], Jing Hua Zhao[1,2], Inga Prokopenko[3,4], Michael Inouye[5], Rachel M Freathy[6,7], Antony P Attwood[5,8], Jacques S Beckmann[9,10], Sonja I Berndt[11], The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial[71], Sven Bergmann[9,12], Amanda J Bennett[3,4], Sheila A Bingham[13], Murielle Bochud[14], Morris Brown[15], Stéphane Cauchi[16], John M Connell[17], Cyrus Cooper[18], George Davey Smith[19], Ian Day[18], Christian Dina[16], Subhajyoti De[20], Emmanouil T Dermitzakis[5], Alex S F Doney[21], Katherine S Elliott[3], Paul Elliott[22,23], David M Evans[3,19], I Sadaf Farooqi[2,24], Philippe Froguel[16,25], Jilur Ghori[5], Christopher J Groves[3,4], Rhian Gwilliam[5], David Hadley[26], Alistair S Hall[27], Andrew T Hattersley[6,7], Johannes Hebebrand[28], Iris M Heid[29,30], KORA[71], Blanca Herrera[3,4], Anke Hinney[28], Sarah E Hunt[5], Marjo-Riitta Jarvelin[22,23,31], Toby Johnson[9,12,14], Jennifer D M Jolley[8], Fredrik Karpe[4], Andrew Keniry[5], Kay-Tee Khaw[32], Robert N Luben[32], Massimo Mangino[33], Jonathan Marchini[34], Wendy L McArdle[35], Ralph McGinnis[5], David Meyre[16], Patricia B Munroe[36], Andrew D Morris[21], Andrew R Ness[37], Matthew J Neville[4], Alexandra C Nica[5], Ken K Ong[1,2], Stephen O'Rahilly[2,24], Katharine R Owen[4], Colin N A Palmer[38], Konstantinos Papadakis[26], Simon Potter[5], Anneli Pouta[31,39], Lu Qi[40], Nurses' Health Study[71], Joshua C Randall[3,4], Nigel W Rayner[3,4], Susan M Ring[35], Manjinder S Sandhu[1,32], André Scherag[41], Matthew A Sims[1,2], Kijoung Song[42], Nicole Soranzo[5], Elizabeth K Speliotes[43,44], Diabetes Genetics Initiative[71], Holly E Syddall[18], Sarah A Teichmann[20], Nicholas J Timpson[3,19], Jonathan H Tobias[45], Manuela Uda[46], The SardiNIA Study[71], Carla I Ganz Vogel[28], Chris Wallace[36], Dawn M Waterworth[42], Michael N Weedon[6,7], The Wellcome Trust Case Control Consortium[72], Cristen J Willer[47], FUSION[71], Vicki L Wraight[2,24], Xin Yuan[42], Eleftheria Zeggini[3], Joel N Hirschhorn[44,48–51], David P Strachan[26], Willem H Ouwehand[8], Mark J Caulfield[36], Nilesh J Samani[33], Timothy M Frayling[6,7], Peter Vollenweider[52], Gerard Waeber[52], Vincent Mooser[42], Panos Deloukas[5], Mark I McCarthy[3,4,73], Nicholas J Wareham[1,2,73] & Inês Barroso[5,73]

Height

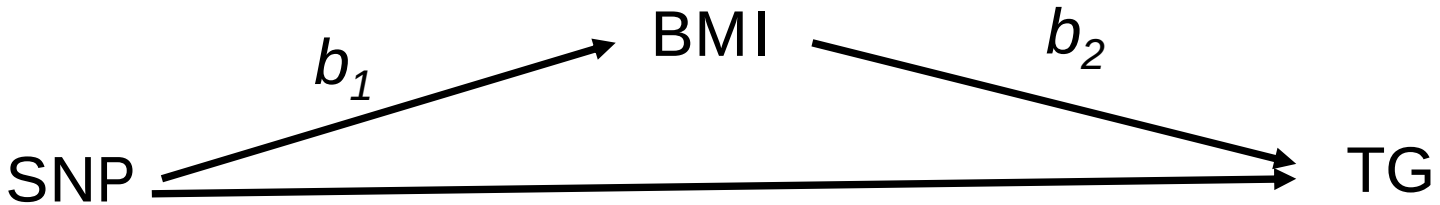
Genome-wide association analysis identifies 20 loci that influence adult height

Michael N Weedon[1,2,23], Hana Lango[1,2,23], Cecilia M Lindgren[3,4], Chris Wallace[5], David M Evans[6], Massimo Mangino[7], Rachel M Freathy[1,2], John R B Perry[1,2], Suzanne Stevens[7], Alistair S Hall[8], Nilesh J Samani[7], Beverly Shields[2], Inga Prokopenko[3,4], Martin Farrall[9], Anna Dominiczak[10], Diabetes Genetics Initiative[21], The Wellcome Trust Case Control Consortium[21], Toby Johnson[11–13], Sven Bergmann[11,12], Jacques S Beckmann[11,14], Peter Vollenweider[15], Dawn M Waterworth[16], Vincent Mooser[16], Colin N A Palmer[17], Andrew D Morris[18], Willem H Ouwehand[19,20], Cambridge GEM Consortium[22], Mark Caulfield[5], Patricia B Munroe[5], Andrew T Hattersley[1,2], Mark I McCarthy[3,4] & Timothy M Frayling[1,2]

Adult height is a model polygenic trait, but there has been limited success in identifying the genes underlying its normal variation. To identify genetic variants influencing adult human height, we used genome-wide association data from 13,665 individuals and genotyped 39 variants in an additional 16,482 samples. We identified 20 variants associated with adult height ($P < 5 \times 10^{-7}$, with 10 reaching $P < 1 \times 10^{-10}$). Combined, the 20 SNPs explain ~3% of height variation, with a ~5 cm difference between the 6.2% of people with 17 or fewer 'tall' alleles compared to the 5.5% with 27 or more 'tall' alleles. The loci we identified implicate genes in Hedgehog signaling (*IHH, HHIP, PTCH1*), extracellular matrix (*EFEMP1, ADAMTSL3, ACAN*) and cancer (*CDK6, HMGA2, DLEU7*) pathways, and provide new insights into human growth and developmental processes. Finally, our results provide insights into the genetic architecture of a classic quantitative trait.

Specific analysis

- Which trait MC4R has effect on?
- Interpretation of mediation
  - Path analysis – shows mainly on BMI and not others
  - Error propagation as appropriate for meta-analysis



As is the case with FTO and T2D, the indirect effect (IE) from MC4R SNP to TG via BMI is $b_1 b_2$, with $SE(IE) \approx b_1 SE(b_2)$

Reflection on the study design

	Study 1 (EPIC-Norfolk subcohort) n=2269		Study 2 (EPIC-Norfolk obese set) n=1009		Study 3 (1958 British birth cohort) n=1375		Study 4 (CoLaus) n=5367		Study 5 (GEMS study) n=1665	
	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value
rs4420638	0·24 (0·04)	$1·9\times10^{-9}$	0·14 (0·06)	0·02	0·25 (0·04)	$2·8\times10^{-9}$	0·05 (0·01)	$6·2\times10^{-12}$	0·04 (0·01)	$5·6\times10^{-3}$
rs599839	−0·15 (0·04)	$5·8\times10^{-5}$	−0·23 (0·06)	$7·6\times10^{-5}$	−0·14 (0·04)	$4·3\times10^{-4}$	−0·04 (0·01)	$1·6\times10^{-7}$	−0·06 (0·01)	$2·0\times10^{-5}$
rs4970834	−0·13 (0·04)	$1·1\times10^{-3}$	−0·18 (0·06)	$5·5\times10^{-3}$	−0·11 (0·04)	0·01	−0·04 (0·01)	$1·9\times10^{-6}$	−0·04 (0·01)	$2·8\times10^{-3}$
rs562338	−0·17 (0·04)	$6·0\times10^{-6}$	−0·11 (0·06)	0·07	−0·18 (0·05)	$1·1\times10^{-4}$	−0·03 (0·01)	$2·7\times10^{-6}$	−0·02 (0·01)	0·18
rs7575840	0·15 (0·03)	$6·3\times10^{-6}$	0·15 (0·05)	$2·4\times10^{-3}$	0·04 (0·04)	0·26	0·03 (0·01)	$1·9\times10^{-6}$	0·02 (0·01)	0·13
rs478442	−0·16 (0·04)	$2·1\times10^{-5}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·6\times10^{-4}$	−0·03 (0·01)	$2·7\times10^{-5}$	−0·02 (0·01)	0·06
rs4591370	−0·17 (0·04)	$7·7\times10^{-6}$	−0·06 (0·06)	0·28	−0·16 (0·04)	$4·2\times10^{-4}$	−0·03 (0·01)	$3·2\times10^{-5}$	−0·02 (0·01)	0·06
rs4560142	−0·16 (0·04)	$1·6\times10^{-5}$	−0·06 (0·06)	0·27	−0·16 (0·04)	$4·2\times10^{-4}$	−0·03 (0·01)	$3·5\times10^{-5}$	−0·03 (0·01)	0·05
rs576203	−0·16 (0·04)	$1·2\times10^{-5}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$3·5\times10^{-5}$	−0·02 (0·01)	0·06
rs506585	−0·16 (0·04)	$1·7\times10^{-5}$	−0·06 (0·06)	0·31	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$4·2\times10^{-5}$	−0·03 (0·01)	0·05
rs488507	−0·14 (0·04)	$1·3\times10^{-4}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·3\times10^{-4}$	−0·03 (0·01)	$3·4\times10^{-5}$	−0·02 (0·01)	0·07
rs538928	−0·16 (0·04)	$5·0\times10^{-5}$	−0·01 (0·06)	0·92	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$3·6\times10^{-5}$	−0·02 (0·01)	0·05
rs10402271	0·04 (0·03)	0·17	0·11 (0·05)	0·02	0·12 (0·04)	$7·5\times10^{-4}$	0·02 (0·01)	$5·2\times10^{-4}$	0·04 (0·01)	$8·3\times10^{-4}$
rs693	−0·12 (0·03)	$1·3\times10^{-4}$	−0·07 (0·05)	0·15	−0·06 (0·03)	0·06	−0·03 (0·01)	$1·0\times10^{-5}$	−0·02 (0·01)	0·16

Webtable 3: Associations between Affymetrix SNPs with a combined p value of $<1·0\times10^{-7}$ and circulating concentrations of LDL cholesterol in independent study populations

Reflection on the analysis

- A characterisation of the EPIC cohort? Maybe!
- an abstract to IGES 2007
  - BMI, or zBMI?
  - QLIM procedure, bivariate with SBP/DBP and HT
- What is the benefit of retrospective method?
- How about staged design?

Allele-coding not a big deal!

Table 1. Allelic coding when the minor allele A is coded as B by alphabetical order

Correct Model	Genotype coding			Coded Model	Genotype coding			Change direction of effect
	A/A	A/B	B/B		A/A	A/B	B/B	
Additive	2	1	0	Additive	0	1	2	Yes
Dominant	1	1	0	Recessive	0	0	1	Yes
Recessive	1	0	0	Dominant	0	1	1	Yes

The bottleneck has been allele-coding, but the inclusion of map information would do away with it.

Our views of data format

- The long and skinny format duplicates information.
- The wide format is familiar but requires specification of SNP names, or use of macro function in the case of SAS.
- The transposed format will be the one to go, with the expectation of a larger data/information base, e.g., the 1000 genome project. Perhaps the major benefit would be the ease with data extraction.
- Or probably it is not a good idea to store the imputed data after all, we can have SAS to grab the data online.

Data extraction for given IDs

- The master file is in long format containing all individuals and all SNPs. Now the following code can be used to extract a subset of individuals with their SNPs on a specific chromosome i.

- proc sql;
- create table genotype as select * from long&i
- where id in (select id from id) &
- rsn in (select rsn from map&i) order by id, rsn;
- quit;

- We can use keep in data step but a simpler grep for ASCII files in transposed format.

Our best practice

- Linux clusters are now ready for comphensive analyses.
- Linux/awk script is light and appears to be more transparent than Perl, Java which is more professional.
- awk proves very useful and can be transformed to Perl. In fact, any statistical package which processes data elements would be less efficient. An example is the transformation of long, wide, transposed format noted earlier.
- They call C/C++ programs such as IMPUTE/SNPTEST.
- SAS is still useful for data preparation, and in a sense less professional than DBMS such as Oracle but enjoys a large user community and has facility for data analysis.
- SAS 9.2 PROTO procedure is yet to be explored.

Linux clusters

- Linux has utilities such as mpirun.
- SAS/Connect can use heterogeneous systems
- It is most useful to use ssh:
  - export GWA=/data/genetics/gwas/3-4-8
  - export RUN=$GWA/id.sh
  - ssh -f c16 "cd $GWA;bash $RUN fatpct 1 12"
  - ssh -f c15 "cd $GWA;bash $RUN fatpct1 1 12"
  - ...
  - ssh -f c09 "cd $GWA;bash $RUN fatpct 13 22"
  - ssh -f c08 "cd $GWA;bash $RUN fatpct1 13 22"
  - ssh c16 "ps x"
  - ssh c15 "ps x"
  - ...
  - ssh c09 "ps x"
  - ssh c08 "ps x"

Meta-analysis (fixed-effects)

```
data test;
        input studyid lor est;
        col=_n_; row=_n_;
        value=est;
Cards;
...  data for 15 studies ...
run;
proc mixed method = ml data=test;
         class studyid;
         model lor = / s cl;
         repeated / group = studyid;
         parms / parmsdata=test eqcons=1 to 15;
run;
```

Meta-analysis (random-effects)

```
proc mixed data=test covtest; /*no specification of 15*/
        class studyid;
        model lor = / s cl outp=predp outpm=predm;
        repeated diag / r;
        random studyid / g gdata = test s v;
        ods output CovParms=cp G=G R=R V=V
                        SolutionF=SF SolutionR=SR;
run;
data predp;
      set predp; pvalue=probnorm(resid/stderrpred);
run;
data predm;
      set predm;pvalue=probnorm(resid/stderrpred);
run;
```

Haplotype analysis

```
%macro wreg(no,model);
proc surveyreg;
        ods output parameterestimates=bmi&no
                        (where=(parameter^="Intercept"));
        cluster id;
        &model / clparm;
        weight p;
run;
data bmi&no;
        model=&no;
        set bmi&no;
run;
%mend wreg;
%wreg(1,model lbmi1=sex age1 h1--h5);
%wreg(2,model lbmi2=sex age2 h1--h5);
```

LD information from hapmap

```
%let url=http://www.hapmap.org/downloads/ld_data/latest;
%let file=ld_chr22_CEU.txt.gz;
filename ldin url "&url/&file" recfm=S;
filename ldout "&file";
data _null_;
      infile ldin;
      file ldout recfm=F;
      input;
      put _infile_;
run;
filename in2 pipe "gzip -dcq &file";
data x.ld;
      infile in2 dlm=' ' lrecl=200 firstobs=2 n=2000;
      format pop $3. rs1 $15. rs2 $15.;
      input pos1 pos2 pop rs1 rs2 dprime r2 lod fbin;
run;
```

Ongoing and further work

- Gene-gene interaction (e.g., height)
- Covariate-SNP (e.g., gender, BMI) interaction
- Gene characterisation
- Family data, e.g., GAW16 and comments by Bodmer and Bonilla:

  "Family studies do not have a significant role in the discovery or analysis of either common or rare disease associated variants, both of which have relatively low penetrances at the individual level."
- Some work on R including gap, kinship, pan, useR!2008 (tutorials and graphical methods).

References

- Bodmer W, Bonilla C. Nat Genet 2008;40:695-701
- EPIC: http://epic.iarc.fr/
- EPIC-Norfolk: http://www.srl.cam.ac.uk/epic
- EPIC-Norfolk GWAS of obesity: http://www.mrc-epid.cam.ac.uk/~jinghua.zhao/software/go.htm
- Long AD et al.. Science 1997; 275:1328
- Loos R et al. Nat Genet 2008
- Prentice RL. Biometrika 1986;73:1-11
- Risch N, Merkangas K (1996) Science 1997;273:1516-7
- Sandhu MS et al. Lancet 2008; 371:483-91
- Weedon MN et al. 2008;40:575-83
- Zhao JH. J Stat Soft 2007;23(8):1-18
- Zhao JH et al. CCIS 2007;2:781-90
- Zhao JH et al. (submitted)

Acknowledgements

EPIC 400: Douglas Easton, Jonathan Morrison

EPIC GWAS: Robert Luben, Ralph McGinnis, Mike Inouye, Simon Potter

Collaboration with many colleagues within MRC Epidemiology Unit: Nick Wareham, Shengxu Li, Ruth Loos, Jian'an Luan, Iain Morrison, Matt Simms